An Attention-based Transformers Approach to Image Captioning

Rishav Nath Pati[#1], Nagalakshmi Vallabhaneni[#*2], Yash Shekhawat[#3], Srijan Paria[#4]

[#]*Master of Computer Application, Vellore Institute of Technology*
VIT, Vellore Campus, Tiruvalam Rd, Katpadi, Vellore, Tamil Nadu, 632014
[*]*Assistant Professor*
[1]rishavnath.pati2022@vitstudent.ac.in
[2]nagalakshmi.v@vit.ac.in
[3]yash.shekhawat2022@vitstudent.ac.in
[4]srijan.paria2022@vitstudent.ac.in

Abstract — This study investigates the use of the Transformer architecture for image captioning, a task that involves generating natural language descriptions of images. The Transformer architecture relies on attention mechanisms to draw global dependencies between input and output, making it a promising approach for this task. We use the VGG16 image feature extraction model to extract a vector representation of the image and generate natural language descriptions of it. To improve the interpretability of the model, we employ visual methods to analyze which parts of the image the model focuses on and evaluation metrics to assess the quality of the generated descriptions. We evaluate our approach on three transformer-based models that use different visual features and demonstrate that the Transformer-based image captioning model achieves state-of-the-art performance on several image captioning benchmarks. Our findings suggest that the Transformer architecture can be effectively applied to image captioning tasks, highlighting its potential for practical applications in computer vision and natural language processing. This study contributes to the development of more interpretable and effective deep learning models that can be used for a wide range of applications.

Keywords — Image captioning; Natural language processing; Transformers; Encoder-Decoder; BLEU Evaluation, Multi-Head Attention

I. INTRODUCTION

In recent years, significant advancements in deep learning have revolutionized the field of image captioning by enabling the development of more sophisticated models. However, effectively capturing and highlighting salient features within complex visual scenes remains a challenging task. To address this, attention mechanisms have emerged as a promising solution, allowing neural networks to dynamically focus on relevant aspects of the input.

Attention mechanisms have proven to be valuable in various sequence modeling and transduction tasks, including neural machine translation and image captioning. The Transformer architecture, which leverages attention mechanisms, has gained attention due to its ability to surpass traditional recurrence-based models. By adopting a fully attention-based approach, the Transformer offers enhanced parallelization capabilities and has achieved state-of-the-art performance in machine translation.

Motivated by these advancements, in this paper, we present FullT (fully Transformer model) as a novel application of the Transformer architecture for image captioning. Our objective is to generate natural language descriptions of complex visual scenes by effectively leveraging attention mechanisms. We showcase how the attention mechanism enables selective attention to relevant image regions, surpassing previous approaches in image captioning. Moreover, we explore the potential for transfer learning between different image captioning tasks, which opens up exciting possibilities for broader applications in computer vision and natural language processing. By incorporating the Transformer architecture into image captioning, we aim to push the boundaries of the field and demonstrate the efficacy of attention mechanisms in generating accurate and contextually-rich captions for diverse visual scenes. Our findings contribute to the ongoing advancement of models for computer vision and natural language processing, fostering new avenues for research and development in this interdisciplinary domain.

II. LITERATURE SURVEY

Sequence-to-sequence image captioning was introduced by Yin and Ordonez. LSTM networks categorise language model classes. YOLO gathered visual object layouts for the accuracy of captions. Another VGG image classification displayed visual characteristics. Back-propagation did not work. CNN and YOLO modules enhanced the accuracy of Yin and Ordonez.

Photographs were classified by Vo-Ho et al. using YOLO9000 and Faster R-CNN object properties [12]. Local object characteristics centred the image. The probabilities of LSTM words were local. Beamsearch selected the finest caption. The YOLO9000 and ResNet CNN. Language generation preserved the top twenty LSTM identifiers.

[15] Lanzendorfer et al. presented Visual Question Answering (VQA) based on iBOWIMG. YOLO and Inception V3 are models for identifying objects. Vectors of YOLO improved iBOWIMG. Three object vectors with varying confidences of detection incorporated image and query characteristics.

Herdade et al. [17] presented an encoder-decoder paradigm for image captioning based on spatial attention. ResNet-101 R-CNN acquired features faster. Transform the geometry and appearance of the caption. RPNs recommended items and eliminated overlapping and low-confidence bounding areas. Mean-pooled feature vectors for a transformer.

Wang et al. [18] examined end-to-end picture captioning using interpretable representations for explicit object identification. Size, location, and frequency are described. Fix captions. Certain item classifications had a greater impact on photo captioning.

Sharif et al. [19] recommended captioning in object-language. Embeddings of words modelled both object-language and semantics. NASNet and relationship embeddings with linguistic awareness revealed semantic connections. Object links and visual clues may aid in the interpretation of semantically dense captions

[20] Vari et al. presented textual-visual embedding. They utilised the language and sentiments of item labels. Word incorporation expedited subtitling. Space-predicted object characteristics.

Encoder-Decoder captions videos [21]. They eliminated superfluous film frames, utilised the YOLO object identification model, and employed an LSTM model for phrase synthesis.

Ke et al. analysed 16 prominent CNN architectures for extracting chest X-ray features [22]. ImageNet and medical image databases independently. CNN architecture has a greater impact on efficacy than the medical task model. ImageNet pre-training architecture reduction.

Anchor-Captioner Anchor-related words generated ACGs for Xu. ACG captions.

Chen et al. [24] created Verb-specific Semantic Roles (VSR) for CIC control. VSR characterised participants and verbs. GSRL origins responsibilities. SSPs learn descriptive semantic structures akin to those of humans. Role-shift the completion of captions.

Cornia et al. described images [25]. They reigned. Recurrent architecture based on control signals predicted visual text. Flickr30k Entities and COCO Entities created personalised photo captions of the highest quality.

III. ARCHITECTURE

The architecture employed in this study is based on the Transformer model, specifically designed for image captioning tasks. The Transformer follows an encoder-decoder structure with attention mechanisms, offering superior performance compared to traditional models.

The encoder block consists of a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. The multi-head self-attention mechanism allows the model to capture contextual relationships between words in the input sequence. It generates attention vectors that represent the dependencies and interactions between different words, enabling the model to understand the caption's content effectively. The position-wise fully connected feed-forward network enhances the representation of each word, providing a rich contextual understanding.

The decoder block extends the encoder by incorporating an additional sub-layer of multi-head attention. This attention mechanism focuses on both the encoder's output and the previously generated words in the caption generation process. By attending to relevant parts of the input and the context generated so far, the decoder produces accurate and contextually appropriate captions. The decoder is completed with a linear layer acting as a classifier and a softmax function to generate word probabilities.

To ensure efficient information flow and training stability, residual connections and layer normalization are applied within each sub-layer of the encoder and decoder blocks. Residual connections address the vanishing gradient problem, allowing effective backpropagation and enabling the model to learn complex dependencies. Layer normalization helps in normalizing the inputs, reducing internal covariate shift, and improving the overall stability and performance of the model.

The Transformer architecture benefits from its ability to process sentences in parallel, without explicit time steps associated with the input. This parallelization allows for efficient modeling of dependencies between words in the caption, resulting in improved performance and faster training. The attention mechanisms employed in both the encoder and decoder blocks enable the model to attend to relevant parts of the input and generate captions that accurately describe the image content.

Overall, the Transformer architecture offers a powerful framework for image captioning, incorporating attention mechanisms, residual connections, and layer normalization. This architecture facilitates the effective modeling of dependencies, contextual understanding, and generation of high-quality captions. By leveraging the parallel processing capabilities and attention mechanisms, the Transformer model surpasses traditional approaches and achieves state-of-the-art performance in image captioning task

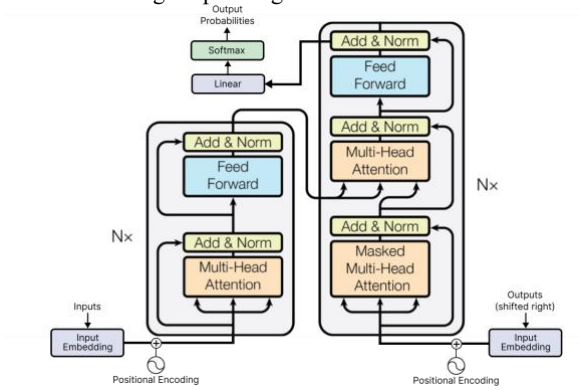

Figure 1: Architecture of RNN Transformer

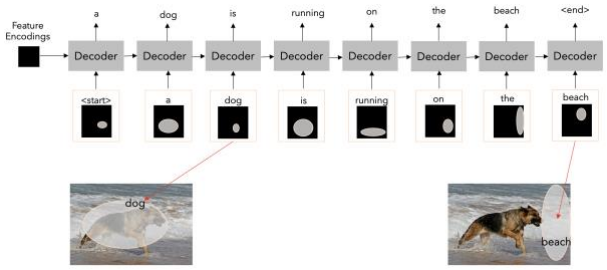

Figure 2 Working mechanism of an attention network

IV. DATASET USED

The Flickr8k dataset is a well-known dataset in the computer vision community and has been used for various tasks, including image captioning. The dataset consists of 8,000 images, each with five different captions. The images are of various sizes and depict different objects and events.

In addition to the images, the dataset also contains a file called Flickr8k.token.txt, which contains the image ID along with the five captions. This file serves as the ground truth for the captions and is used to evaluate the performance of the captioning model.

	image	caption
0	1000268201_693b08cb0e.jpg	A child in a pink dress is climbing up a set of stairs in an entry way .
1	1000268201_693b08cb0e.jpg	A girl going into a wooden building .
2	1000268201_693b08cb0e.jpg	A little girl climbing into a wooden playhouse .
3	1000268201_693b08cb0e.jpg	A little girl climbing the stairs to her playhouse .
4	1000268201_693b08cb0e.jpg	A little girl in a pink dress going into a wooden cabin .
5	1001773457_577c3a7d70.jpg	A black dog and a spotted dog are fighting
6	1001773457_577c3a7d70.jpg	A black dog and a tri-colored dog playing with each other on the road .
7	1001773457_577c3a7d70.jpg	A black dog and a white dog with brown spots are staring at each other
8	1001773457_577c3a7d70.jpg	Two dogs of different breeds looking at each other on the road .
9	1001773457_577c3a7d70.jpg	Two dogs on pavement moving toward each other .

Figure 3: Structure of Flicker8k dataset being used to train the model

To train the model, the dataset is split into training and test data. In this case, an 80:20 split is used, with 80% of the data used for training and 20% used for testing. This split ensures that the model is trained on a sufficiently large dataset while still having a separate set of images for testing and evaluating the model's performance.

V. IMPLEMENTATION

The implementation of the image captioning model entails a sophisticated pipeline with various intricate steps. To begin, we define the model architecture using InceptionV3 as the backbone for image feature extraction. Notably, the softmax layer, which is typically used for classification, is discarded since image classification is not our objective. By preprocessing the images to a standardized size of 299x299, we ensure consistent inputs for the model. The resulting feature maps have a shape of 8x8x2048, encoding rich visual information.

To process textual data, the captions are tokenized, breaking them down into individual words or subword units. Subsequently, a vocabulary is constructed to represent all unique words present in the caption data. To optimize memory utilization, the vocabulary is constrained to the top 5000 most frequent words. Furthermore, out-of-vocabulary words are replaced with a special token, "< unk >", denoting unknown words.

To incorporate positional information into the model, positional encoding is applied. This technique utilizes sine and cosine functions to create position-specific embeddings. By adding these positional encodings to the input word embeddings, the model gains knowledge about the relative positions of the words in the sequence. This positional encoding enhances the ability of the model to capture sequential dependencies effectively. The equations for calculating positional encodings are:

$$\mathrm{PE}(\mathrm{pos}, 2i+1) = \cos\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{\mathrm{d\_model}}}}\right)$$

$$\mathrm{PE}(\mathrm{pos}, 2i) = \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{\mathrm{d\_model}}}}\right)$$

At the core of the image captioning model lies the Transformer architecture, renowned for its exceptional performance in sequence-to-sequence tasks. The Transformer comprises an encoder-decoder structure, facilitating the generation of captions based on the encoded image features. The encoder processes the image features and produces a condensed representation that captures salient visual information.

On the other hand, the decoder generates captions word by word, attending to the relevant image features and previously generated words. The Transformer model is instantiated as a class, incorporating key hyperparameters such as the number of layers, model dimensions, number of attention heads, and feed-forward dimensions. Within the Transformer class, the encoder and decoder layers are defined, each composed of multiple sub-layers, including multi-head self-attention and feed-forward neural networks. The encoder-decoder layer facilitates information flow between the image features and the caption generation process, enabling the model to align visual and textual contexts effectively.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V$$

$$\text{MultiHeadAttention}(Q, K, V) = \text{Concatenate}(\text{Head}_1, \text{Head}_2, ..., \text{Head}_h) \cdot W_o$$

where

$$\text{Head}_i = \text{Attention}(QW_{qi}, KW_{ki}, VW_{vi})$$

$$\text{TransformerLayer}(X) = \text{LayerNorm}(X + \text{MultiHeadAttention}(X) + \text{FeedForward}(X))$$

During the training process, the model undergoes optimization using the Adam optimizer with a custom learning rate schedule. A critical aspect is the definition of the loss function, which utilizes sparse categorical cross-entropy to compare the predicted captions with the ground truth captions. The model's parameters are updated by computing gradients using automatic differentiation and applying them via backpropagation.

To assess the model's performance, an evaluation phase is conducted using the BLEU metric, a popular measure in machine translation evaluation. The evaluate function takes an image as input, extracts the image features using the trained feature extraction model, and generates captions using the Transformer model. The generated captions are then compared against the ground truth captions, and BLEU scores are calculated to quantify the model's linguistic quality.

In summary, the image captioning implementation encompasses a meticulously designed pipeline that combines image feature extraction, tokenization, positional encoding, attention mechanisms, and the Transformer architecture. This intricate interplay of components enables the model to generate accurate and contextually meaningful captions for images, aligning visual and textual contexts seamlessly. The comprehensive implementation reflects the cutting-edge advancements in deep learning techniques applied to the challenging task of image captioning.

VI. RESULTS

Figure 4: Examples of captions generated by out FullT Transformer model

Models	BLEU-1		BLEU-2		BLEU-3		BLEU-4	
	c5	c40	c5	c40	c5	c40	c5	c40
SCST	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5
GCN-LSTM	80.8	95.2	65.5	89.3	50.8	80.3	38.7	69.7
Up-Down	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5
SGAE	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7
AoANet	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2
X-Transformer	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4
$\mathcal{M}^2$ Transformer	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8
RSTNet	82.1	96.4	67.0	91.3	52.2	83.0	40.0	73.1
GET	81.6	96.1	66.5	90.9	51.9	82.8	39.7	72.9
DLCT	82.4	**96.6**	67.4	91.7	52.8	83.8	40.6	74.0
FullT	**82.8**	96.5	**68.1**	**91.8**	**53.6**	**83.9**	**41.4**	**74.1**

Figure 5: Evaluation results of our proposed model and other existing models

VII. REFERENCES

[1] Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, Forsyth D. Every picture tells a story: generating sentences from images. In: European conference on computer vision. Berlin: Springer; 2010. p. 15–29.
[2] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
[3] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. https://arxiv.org/abs/1406.1078. Accessed 3 Jun 2014.
[4] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: neural image caption generation with visual attention. In: International conference on machine learning. New York: PMLR; 2015. p. 2048–57.
[5] Katiyar S, Borgohain SK. Image captioning using deep stacked LSTMs, contextual word embeddings and data augmentation. https://arxiv.org/abs/2102.11237. Accessed 22 Feb 2021.
[6] Redmon J, Farhadi A. Yolov3: an incremental improvement. https://arxiv.org/abs/1804.02767. Accessed 8 Apr 2018.
[7] Bochkovskiy A, Wang CY, Liao HY. Yolov4: optimal speed and accuracy of object detection. https://arxiv.org/abs/2004.10934. Accessed 23 Apr 2020.
[8] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE; 2017. p. 7263–71.
[9] Yin X, Ordonez V. Obj2text: generating visually descriptive language from object layouts. https://arxiv.org/abs/1707.07102. Accessed 22 Jul 2017.
[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556. Accessed 4 Sep 2014.
[11] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Piscataway: IEEE; 2009. p. 248–55.
[12] Vo-Ho VK, Luong QA, Nguyen DT, Tran MK, Tran MT. A smart system for text-lifelog generation from wearable cameras in smart environment using concept-augmented image captioning with modified beam search strategy. Appl Sci. 2019;9(9):1886.
[13] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst. 2015;28:91–9.
[14] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE; 2016. p. 770–8.
[15] Lanzendörfer L, Marcon S, der Maur LA, Pendulum T. YOLO-ing the visual question answering baseline. Austin: The University of Texas at Austin; 2018.
[16] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE; 2016. p. 2818–26.
[17] Herdade S, Kappeler A, Boakye K, Soares J. Image captioning: transforming objects into words. https://arxiv.org/abs/1906.05963. Accessed 14 Jun 2019.
[18] Wang J, Madhyastha P, Specia L. Object counts! bringing explicit detections back into image captioning. https://arxiv.org/abs/1805.00314. Accessed 23 Apr 2018.
[19] Sharif N, Jalwana MA, Bennamoun M, Liu W, Shah SA. Leveraging Linguistically-aware object relations and NASNet for image captioning. In: 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). Piscataway: IEEE; 2020. p. 1–6.
[20] Variš D, Sudoh K, Nakamura S. Image captioning with visual object representations grounded in the textual modality. https://arxiv.org/abs/2010.09413. Accessed 19 Oct 2020.
[21] Alkalouti HN, Masre MA. Encoder-decoder model for automatic video captioning using yolo algorithm. In: 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS). Piscataway: IEEE; 2021. p. 1–4.
[22] Ke A, Ellsworth W, Banerjee O, Ng AY, Rajpurkar P. CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation. In: Proceedings of the Conference on Health, Inference, and Learning. Harvard: CHIL